A Dataset for Learning University STEM Courses at Scale and Generating Questions at a Human Level

Iddo Drori[1, 4, 5], Sarah Zhang[1], Zad Chin[2], Reece Shuttleworth[1], Albert Lu[1], Linda Chen[1], Bereket Birbo[1], Michele He[1], Pedro Lantigua[1], Sunny Tran[1], Gregory Hunter[4], Bo Feng[4], Newman Cheng[4], Roman Wang[4], Yann Hicke[3], Saisamrit Surbehera[4], Arvind Raghavan[4], Alexander Siemenn[1], Nikhil Singh[1], Jayson Lynch[6], Avi Shporer[1], Nakul Verma[4], Tonio Buonassisi[1], Armando Solar-Lezama[1]

[1]Massachusetts Institute of Technology,
[2]Harvard University,
[3]Cornell University,
[4]Columbia University,
[5]Boston University,
[6]University of Waterloo,
{idrori@mit.edu, idrori@cs.columbia.edu, idrori@bu.edu, sazhang@mit.edu, zadchin@college.harvard.edu, rshuttle@mit.edu, albert03@mit.edu, linda55@mit.edu, bereketb@mit.edu, mjhe@mit.edu, lantigua@mit.edu, sunnyt@mit.edu, geh2129@columbia.edu, bf2477@columbia.edu, nc2893@columbia.edu, rzw2002@columbia.edu, ylh8@cornell.edu, ss6365@columbia.edu, ar4284@columbia.edu, asiemenn@mit.edu, nsingh1@mit.edu, jayson.lynch@waterloo.ca, jayson.lynch@waterloo.ca, shporer@mit.edu, verma@cs.columbia.edu, buonassi@mit.edu, asolar@csail.mit.edu}

Dataset

https://github.com/idrori/stemQ

ID	University	Department	Course	Number
1	MIT	Mechanical Engineering	Hydrodynamics	2.016
2	MIT	Mechanical Engineering	Nonlinear Dynamics I: Chaos	2.050J
3	MIT	Mechanical Engineering	Information & Entropy	2.110J
4	MIT	Mechanical Engineering	Marine Power and Propulsion	2.611
5	MIT	Materials Science and Engineering	Fundamentals of Materials Science	3.012
6	MIT	Materials Science and Engineering	Math for Materials Scientists & Engineers	3.016
7	MIT	Materials Science and Engineering	Introduction to Solid-State Chemistry	3.091
8	MIT	Chemistry	Principles of Chemical Science	5.111
9	MIT	Electrical Engineering & Computer Science	Signal Processing	6.003
10	MIT	Electrical Engineering & Computer Science	Introduction to Machine Learning	6.036
11	MIT	Electrical Engineering & Computer Science	Introduction to Probability	6.041
12	MIT	Physics	Quantum Physics	8.04
13	MIT	Physics	Introduction to Astronomy	8.282
14	MIT	Earth, Atmospheric & Planetary Sciences	Geobiology	12.007
15	MIT	Economics	Principles of Microeconomics	14.01
16	MIT	Aeronautics and Astronautics	Unified Engineering 1–2	16.01–02
17	MIT	Aeronautics and Astronautics	Unified Engineering 3–4	16.03–04
18	MIT	Mathematics	Probability and Random Variables	18.600
19	MIT	Mathematics	Theory of Numbers	18.781
20	MIT	Biological Engineering	Systems Microbiology	20.106J
21	MIT	Institute for Data, Systems & Society	Statistical Thinking & Data Analysis	IDS.013J
22	Brown	Mathematics	Intermediate Calculus	MATH0180
23	Cornell	Computer Science	Computer Architecture	CS4420
24	Harvard	Statistics	Probability	STATS110
25	Princeton	Mathematics	Calculus II	MATH104
26	UPenn	Mathematics	Calculus	MATH110
27	Yale	Mathematics	Fundamentals of Physics	PHYS200

Table 1: A new dataset of questions and solutions from STEM courses by university and department: 27 courses across a dozen departments in seven universities. We curate a dataset and generate new questions for each course.

Datasets as Benchmarks

We would like datasets to serve as benchmarks for measuring performance of large language models

Problem: once the dataset of questions and answers is available online the large language model uses the dataset for training.

Solution: provide private access to new datasets

Mathematical Abilities of Large Language Models

Large language models are trained using optimization objectives such as filling in missing words and if two sentences follow each other.

Problem: how can such models be used to solve STEM problems requiring precision using arbitrary values?

Solution: programs, few-shot learning, chain of thought.

Method

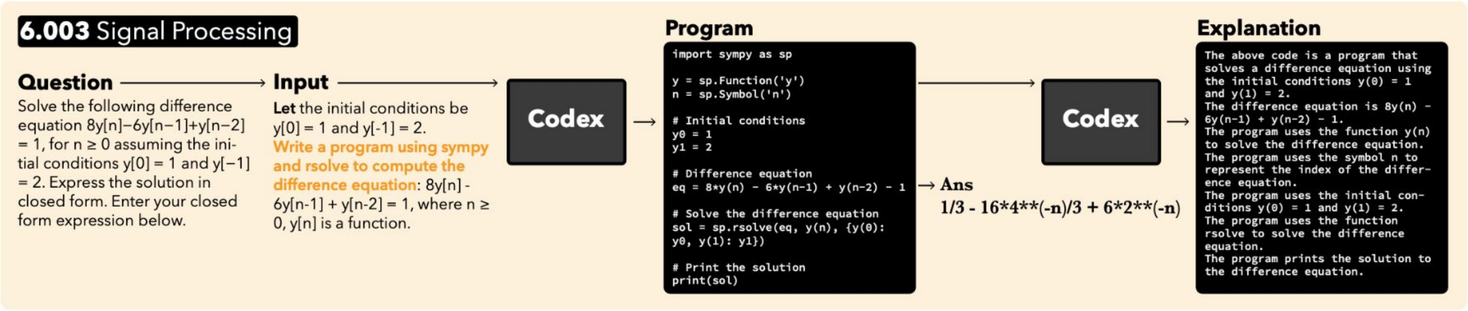


Figure 2: MIT 6.003 Signal Processing workflow: The question is solved as is and the prompt adds programming context to use symbolic math sympy package to produce code snippets that generate answers in form of a symbolic mathematical equation.

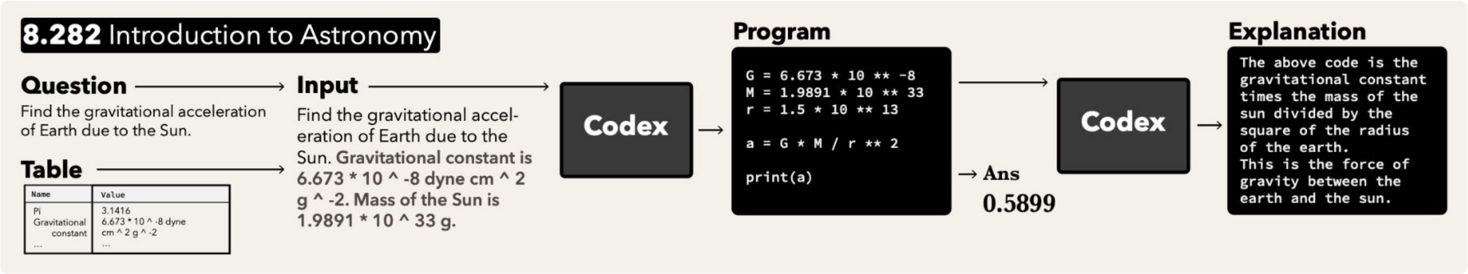


Figure 3: MIT 8.282 Introduction to Astronomy workflow: In this course, Codex often requires context about physical constants. This question involves the gravitational constant (G), mass of the Sun (M), and the distance between the Earth and the Sun (r; this is the definition of one Astronomical Unit).

Few-Shot Learning

Asking a large language model STEM questions

Problem: is analogous to asking a human a question without learning the subject

Solution: few-shot learning allows to provide other question-answer examples before the question

Curriculum Design

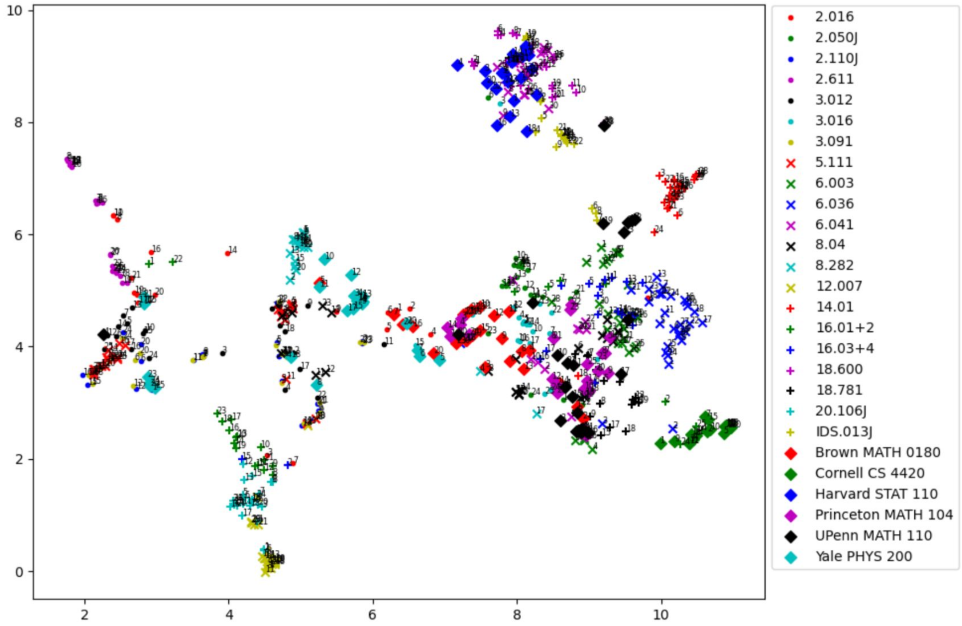


Figure 1: Visualization of embeddings of course questions: We embed the course questions into a 2,048-dimensional space using OpenAI's *text-similarity-babbage-001* embedding, which captures semantic similarity between texts. We then use uniform manifold approximation and projection to reduce the dimensionality to two, observing distinct clusters based on course topics. On the top right, we see a cluster of questions from probability and statistics courses: MIT's 6.041 Introduction to Probability, 18.600 Probability and Random Variables, MIT's IDS013J Statistical Thinking and Data Analysis, and Harvard's STAT 110. On the left side, we see a cluster of the questions from Mechanical Engineering courses: MIT's 2.611, 2.016, and 2.110J. Next, we see a cluster of questions from Chemistry and Materials Science and Engineering on the left. On the bottom left, we see a cluster of the questions from Aeronautics and Astronautics: 16.01, 16.02, 16.03, and 16.04. Below is a cluster of questions from MIT's 20.106J System Microbiology and 12.007 Geobiology. In the center, we see a cluster of questions from Yale PHYS 200, MIT's Quantum Physics, and Introduction to Astronomy, all related to Physics. On the right, we see a cluster of questions from MIT EECS courses. A cluster of questions from math courses appears in the center between EECS and Physics.

Curriculum Design

Problem: what questions, topics, and classes help answer other questions, understand other topics, and are prerequisites for other classes?

Solution: embed questions in low-dimensional space and show the relationships between questions, topics, and classes providing insight into course prerequisites and curriculum design based on data.

Generating Questions

ID	Course	Method	Question
4	Nonlinear Dynamics I: Chaos	Human	Find all the fixed points of the flow $\dot{x} = \sin(x)$
		Machine	Approximate the value of r at which the logistic map has a superstable 4-cycle. Please give a numerical approximation that is accurate to at least four places after the decimal point.
5	Fundamentals of Materials Science	Human	In discussing molecular rotation, the quantum number J is used rather than l. Using the Boltzmann distribution, calculate $\frac{n_J}{n_0}$ for $H_{35}Cl$ for $J = 0$, 5, 10, and 20 at $T = 1025$ K.
		Machine	A sample of an ideal gas is heated reversibly and adiabatically from an initial temperature of 300 K to a final temperature of 600 K. If the initial volume of the sample is 1.00 L, what is the final volume of the sample?
8	Principles of Chemical Science	Human	Using the table of mean bond enthalpies provided, predict the bond enthalpy (in kJ/mol) for the CO bond marked with an arrow in the molecule below. (Bond enthalpeis: C-H - 412 kJ/mol, C-C - 348 kJ/mol, C=C - 612 kJ/mol, C-O - 320 kJ/mol, C=O - 743 kJ/mol).
		Machine	10.0 mL sample of 0.20 M HNO_2 (aq) solution is titrated with 0.10 M NaOH (aq). (Ka of HNO_2 is 4.3×10^{-4}). Calculate the pH of the solution when 10.0 mL of 0.10 M NaOH has been added.
9	Signal Processing	Human	Determine the Laplace transforms (including the regions of convergence) of the following signal: $x(t) = (1 - (1 - t) * \exp(-3 * t))u(t)$.
		Machine	Find the inverse Fourier transforms of the following signal: $X(f) = (\frac{\pi}{2}) * \text{sinc}(\pi * f)$
13	Introduction to Astronomy	Human	Measurements of the radial recession velocity of five galaxies in a cluster give velocities of 9700, 8600, 8200, 8500, and 10000 km s^{-1}. What is the distance to the cluster if the Hubble parameter is $H_0 = 72$ km s^{-1} Mpc^{-1} ? Hint: Use the Hubble law for the average velocity of the members in the cluster.
		Machine	A star has a luminosity of $10^4 L_\odot$ and a temperature of 10^4 K. Find the star's radius in units of the Sun's radius.
14	Geobiology	Human	Determine the kind of nonmarine sedimentary deposits that reflects arid environmental conditions.
		Machine	Determine the term 'Precambrian shield' and the place where it is present in North America
15	Principles of Microeconomics	Human	Chloe consumes only books (x) and video games (y). Her preferences can be represented by the following utility function: $U(x, y) = x * (y^2)$. Calculate the Marginal Rate of Substitution (at an arbitrary bundle (x, y)).
		Machine	Suppose the demand for apples is $Q_D = 550 - 50 * P$ and the industry supply curve is $Q_S = -12.5 + 62.5 * P$. Calculate the equilibrium price and quantity.
16	Unified Engineering 1-2	Human	Define a thermoplastic and a thermoset.
		Machine	What is the difference between an isothermal and an isentropic process?
19	Theory of Numbers	Human	Tabulate the number of primes less than x, for $x = 10000, 20000, \ldots, 100000$. Also tabulate the number of primes less than x and of the form $4k + 1$, and the number of the form $4k + 3$.
		Machine	Find the smallest integer $n > 1$ such that n^2 divides the factorial of n.
20	Systems Microbiology	Human	What does proton motive force mean, and why is it important in biology?
		Machine	Describe the difference between a batch culture and a continuous culture.
22	Intermediate Calculus	Human	Find the value of $\frac{dz}{dx}$ at the point (1, 1, 1) if the equation $xy + z^3x - 2yz = 0$ defines z as a function of the two independent variables x and y and the partial derivative exists.
		Machine	Find the surface area of the portion of the paraboloid $z = 4 - x^2 - y^2$ that lies above the xy-plane.
23	Computer Architecture	Human	What is 01100110 XOR 00111011 in binary?
		Machine	The CPU runs an operating system kernel. A user process occupies the bottom half of the 32-bit address space (i.e., the lower addresses), while the kernel occupies the top half of the same address space (i.e., the higher address) What is the address of the first byte of the kernel?
24	Probability	Human	You have a group of couples that decide to have children until they have their first girl, after which they stop having children. What is the expected gender ratio of the children that are born? What is the expected number of children each couple will have?
		Machine	A fair coin is tossed repeatedly until a head is followed by a tail. What is the expected number of coin tosses?

Table 2: A subset of the human- and machine-generated questions for various courses. The table of questions for all of the 27 STEM courses is found in the Supplementary Material.

Generating Questions

Large language models generate new questions indistinguishable from human-written questions

Problem: are the answers correct?

Solution:
Automatic checkers for several types of questions.
Evaluate students by asking them to evaluate if answers are correct.
Training a classifier predicting if model can answer question.

Id	Approach	Accuracy
1	Specialized model trained on astronomy	92%
2	Generalized Codex (Chen et al. 2021)	67%
3	GPT-3 with CoT (Kojima et al. 2022a)	37.5%
4	GPT-3 (Brown et al. 2020)	30%
5	GPT-3 with CoT (Kojima et al. 2022a)	27%
6	GPT-3 (Brown et al. 2020)	15%
7	Jurassic-1 (Lieber et al. 2021)	10%
8	Wolfram Alpha (Wolfram 2021)	0%
9	Specialized model (Tran et al. 2021)	0%

Table 3: Comparison of accuracy on Introduction to Astronomy course questions. The specialized approach trained on Astronomy achieves 92% accuracy. The generalized approach of writing programs to solve the questions, and synthesizing the programs using OpenAI Codex, achieves 67%. GPT-3 (text-davinci-003) with chain-of-thought (CoT) prompting achieves 37.5%, and GPT-3 (text-davinci-003) without chain-of-thought prompting 30%. GPT-3 (text-davinci-002) with chain-of-thought (CoT) prompting achieves 27% and GPT-3 (text-davinci-002) without chain-of-thought prompting 15%. Jurassic-1 achieves 10%. Wolfram Alpha and a specialized model trained on a different course completely fail.

Performance and Scale

Large language models are generalist and scalable.

Performance improves using

- Few-shot learning
- Chain of thought
- Program synthesis
- Self-error correction

Conclusions

Language models train on online data so for datasets to become benchmarks they should be released privately.

When asking large language models STEM questions it makes sense to give them at least same learning methods available to humans, so they perform at a human level.
Learn from previous examples: few-shot learning
Use chain of thought and program synthesis
Provide multiple attempts by self error correction.

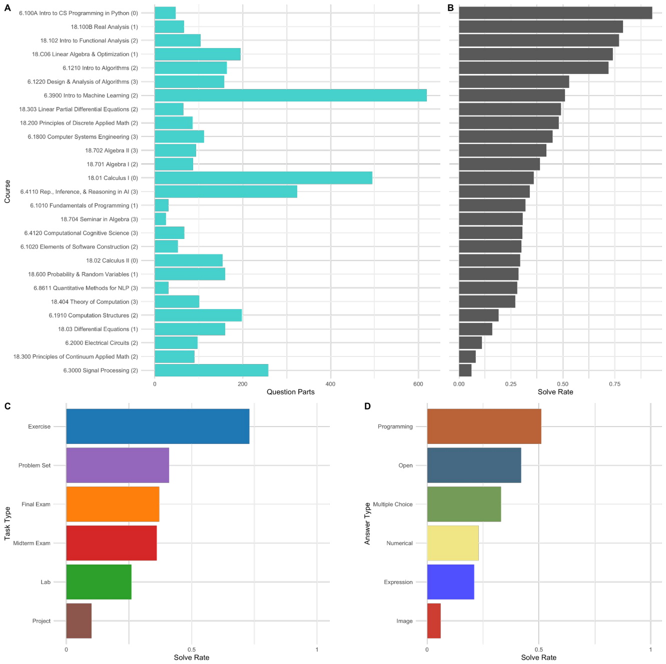
We generate hundreds of questions automatically from other questions, class notes, and books.

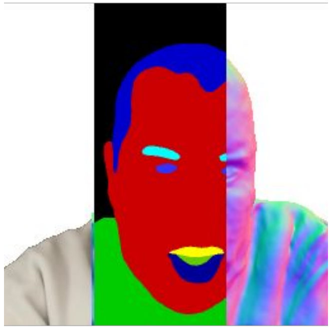
What Now

ChatGPT solves “only” a third of the MIT Mathematics and EECS curriculum. Using our approaches solves the entire curriculum; methods for handling images and proofs; RL with LM inside.

Curriculum design based on data.

Photorealistic speaking avatars delivering machine generated content.





10 Most Recent Related Publications

Prospects and perils of writing books in Mathematics and Computer Science using AI
Iddo Drori, Sarah J. Zhang, Sage Simhon, Yann Hicke, Zad Chin, Keith Tyser, Harsh Sharma, Kirsi Rajagopal, Alice Zhang, Annie Wang, Eugenia Feng, Nikhil Singh, Lauren Cowles, Tonio Buonassisi, Madeleine Udell, Gilbert Strang, Armando Solar-Lezama
In progress

Automatically fulfilling the MIT Mathematics and EECS graduation requirements at a human level by self error correction and few-shot learning
Iddo Drori, Sarah J. Zhang, Sage Simhon, Keith Tyser, Sarah Zhang, Reece Shuttleworth, Pedro Lantigua, Arvind Raghavan, Zad Chin, Saisamrit Surbehera, Leonard Tang, Yann Hicke, Avi Shporer, Nakul Verma, Tonio Buonassisi, Gilbert Strang, Armando Solar-Lezama
Under review

ChatMIT: A dataset for graduating from MIT Mathematics & EECS, achieving human solve rates, curriculum analysis, and generating class questions
Sarah J. Zhang, Sage Simhon, Yann Hick, Zad Chin, Annie Wang, Kirsi Rajagopal, Alice Zhang, Eugnia Feng, Kieth Tyser, Harsh Sharma, Tonio Buonassisi, Armando Solar-Lezama, Iddo Drori
Under review

Text to graphics by program synthesis with error correction on precise, procedural, and simulation tasks
Arvind Raghavan, Zad Chin, Alexander E. Siemenn, Vitali Petsiuk, Saisamrit Surbehera, Yann Hicke, Edward Chien, Ori Kerret, Tonio Buonassisi, Kate Saenko, Armando Solar-Lezama, Iddo Drori
Under review

From human days to machine seconds: Automatically answering and generating machine learning final exams
Sarah J. Zhang, Keith Tyser, Sage Simhon, Sarah Zhang, Reece Shuttleworth, Zad Chin, Pedro Lantigua, Saisamrit Surbehera, Gregory Hunter, Derek Austin, Yann Hicke, Leonard Tang, Sathwik Karnik, Darnell Granberry, Iddo Drori
Under review

A dataset for learning university STEM courses at scale and generating questions at a human level
Iddo Drori, Sarah Zhang, Zad Chin, Reece Shuttleworth, Albert Lu, Linda Chen, Bereket Birbo, Michele He, Pedro Lantigua, Sunny Tran, Gregory Hunter, Bo Feng, Newman Cheng, Roman Wang, Yann Hicke, Saisamrit Surbehera, Arvind Raghavan, Alexander Siemenn, Nikhil Singh, Jayson Lynch, Avi Shporer, Nakul Verma, Tonio Buonassisi, Armando Solar-Lezama
Educational Advances in Artificial Intelligence (EAAI), 2023.

Human evaluation of text-to-image models on a multi-task benchmark
Vitali Petsiuk, Alexander Siemenn, Saisamrit Surbehera, Zad Chin, Kieth Tyser, Gregory Hunter, Arvind Raghavan, Yann Hicke, Bryan Plummer, Ori Kerret, Tonio Buonassisi, Kate Saenko, Armando Solar-Lezama, Iddo Drori
NeurIPS Workshop on Human Evaluation of Generative Models, 2022.

A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level
Iddo Drori, Sarah Zhang, Reece Shuttleworth, Leonard Tang, Albert Lu, Elizabeth Ke, Kevin Liu, Linda Chen, Sunny Tran, Newman Cheng, Roman Wang, Nikhil Singh, Taylor L. Patti, Jayson Lynch, Avi Shporer, Nakul Verma, Eugene Wu, Gilbert Strang
Proceedings of the National Academy of Sciences (**PNAS**), 119(32), 2022.

Solving Probability and Statistics problems by probabilistic program synthesis at human level and predicting solvability
Leonard Tang, Elizabeth Ke, Nikhil Singh, Bo Feng, Derek Austin, Nakul Verma, Iddo Drori
International Conference on Artificial Intelligence in Education (AIED), 2022.

Solving machine learning problems
Sunny Tran, Pranav Krishna, Ishan Pakuwal, Prabhakar Kafle, Nikhil Singh, Jayson Lynch, Iddo Drori
Asian Conference on Machine Learning (ACML), 2021.
Best paper award winner

Spring 2023 Team

Iddo Drori, MIT, Columbia University, Boston University
Yann Hicke, Cornell University
Sarah Zhang, MIT
Sage Simhon, MIT
Zad Chin, Harvard
Annie Wang, MIT
Alice Zhang, MIT
Eugenia Feng, MIT
Samuel Florin, MIT
Harsh Sharma, Boston University
Keith Tyser, Boston University
Andrei Marginean, MIT
Saisamrit Surbehera, Columbia University
Nikhil Singh, MIT
Leonard Tang, Harvard
Lauren Cowles, Cambridge University Press
Gilbert Strang, MIT
Tonio Buonasisi, MIT
Madeleine Udell, Stanford
Armando Solar Lezama, MIT